Segmentation strategies to face morphology challenges in Brazilian-Portuguese/English statistical machine translation and its integration in cross-language information retrieval

Marta R. Costa-jussà[1]

Computer Science Department
Institute of Mathematics and Statistics
University of São Paulo, Brazil
Rua do Matão 1010, São Paulo, SP 05508-090
{martarcj}@ime.usp.br

Abstract. The use of morphology is particularly interesting in the context of statistical machine translation in order to reduce data sparseness and compensate any lack of training corpus. In this work, we propose several approaches to introduce morphology knowledge into a standard phrase-based machine translation system. We provide word segmentation using two different tools (COGROO and MORFESSOR) which allow to reduce the vocabulary and data sparseness. Then, we add to these segmentations the morphological information of a POS language model. We combine all these approaches using a Minimum Bayes Risk strategy. Experiments show significant improvements from the enhanced system over the baseline system on Brazilian Portuguese/English language pair. Finally, we report a case study about the impact of enhancing the statistical machine translation system with morphology in a cross-language application system such as ONAIR which allows users to look for information in video fragments through queries in natural language.

Keywords. Morphology, Factored-based Machine Translation, Cross-language Information Retrieval

Estrategias de segmentación para afrontar los retos morfológicos en un sistema de traducción automática estadística y su integración en un sistema de búsqueda de información crosslingüe.

Resumen. La segmentación basada en morfología es particularmente interesante en el contexto de la traducción automática estadística para reducir la dispersión de los datos y compensar la falta de los mismos. En este trabajo, proponemos diferentes aproximaciones para incorporar conocimiento morfológico en un sistema estándar de traducción automática estadística basado en segmentos. Propocionamos una segmentación en palabras usando dos herramientas diferentes (COGROO y MORFESSOR) que permiten precisamente reducir la dispersión de los datos. Después, añadimos información morfológica al modelo de lenguaje. Combinamos estas aproximaciones usando una estrategia para minimizar el riesgo de Bayes. Los experimentos demuestran mejoras significativas del sistema mejorado respecto al sistema de referencia en el par Brasileño/Inglés. Finalmente, reportamos un caso de estudio sobre el impacto de mejorar el sistema de traducción estadística con morfología en una aplicación como ONAIR que permite a los usuarios buscar información en fragmentos audiovisuales usando consultas en lenguaje natural.

Palabras clave. Morfología, Traducción Automática Estadística Factorizada, Búsqueda de Información Crosslingüe

1 Introduction

The information society is generating a vast quantity of multilingual information, which strongly motivates the use of machine translation (MT) systems. That is why nowadays research in the area of MT is very active. Among the different MT techniques, there are the rule-based techniques [15] and the corpus-based techniques such as statistical [24] or example-based [39]. The former requires a very strong knowledge of the pair of languages involved in the translation, whereas the latter requires a certain amount of bilingual corpora as training data in order to achieve competitive results.

In particular, in this paper, we are using the standard phrase-based statistical machine translation (SMT) approach [24]. Such SMT system uses models, which need a large amount of data to train, so as to estimate the probabilities of the language model and translation model and ensure that the models can accurately estimate probabilities for a majority of forms. Without enough data available, the main issue is data sparseness. This sparseness is even higher in the case of languages with a rich morphology. Recent advances in statistical-based approaches try to introduce linguistic knowledge in order to complement the lack of bilingual corpora, which may never be sufficient.

The use of morphology is particularly interesting in the sense that if we have seen the form *house* in our training corpus, we should be able to translate the corresponding plural form *houses* as well, even though it has never been seen in the training corpus. In this sense, we propose to follow an approach based on morpheme segmentation, being a morpheme the smallest semantic unit from a language. We use special tools for providing this segmentation. Additionally, we experiment with the introduction of additional language models into the standard phrase-based approach which make use of morphological information. We report consistent improvements both in an in-domain and out-domain evaluation sets.

Additionally, after our morphology study, we report a case study in Cross-language Information Retrieval (CLIR). The main objective is to evaluate the influence of the improvement in morphology in a real application. The application we are focusing on is in the context of looking for information in digital videos. Generally, the user can save time, by avoiding to browse through hours of video. Additionally, these videos may be in a foreign language. Although the user may be able to understand the foreign language, he/she may not be able to formulate a query. The ONAIR (Ontology-Aided Information Retrieval) system[2], which started in 2003, intended to allow users to look for information in video fragments through queries in natural language. We study a multilingual extension of this application by further enhancing previous works [10].

The rest of the paper is organised as follows. Section 2 includes a brief related work (without aiming at completeness) in using morphology knowledge for improving MT. The following section reports the phrase-based SMT approach. Section 4 explains in detail the two methods that we propose to integrate morphology knowledge to enhance the phrase-based approach. Then, Section 5 contains the description of experiments performed to evaluate the quality of our proposed morphology integration. Also, we include analysis and discussion of the results. Section 6 describes our case study. Finally, Section 7 concludes with the most relevant contributions of this work.

[2]http://www.ime.usp.br/ rmcobe/onair/

2 Related Work

The challenges raised when translating from or into richer morphology languages are well known and are being continuously studied in the context of SMT. Morphology is the study of the structure of a given language's morphemes, which are the primitive units of syntax, the smallest individually meaningful elements in the utterances of a language. The most important morpheme is the stem, which is the root of the word. The affixes provide additional meaning to the main concept provided by the stem [18].

Morphologically-rich languages have many different surface forms for the same stem. This leads to rapid vocabulary growth, as various prefixes and suffixes can combine with stems in a large number of possible combinations and worse language model probability estimation. There are more singletons (forms occurring just once in the data), and less occurrences over all distinct words. The problem of morphology sparsity becomes even more crucial when addressing translations out-domain. Under that scenario, there is a high presence of previously unseen inflected forms even though their stem could have been learned with the training material. The sparsity due to morphology can be reduced by incorporating morphological information into the SMT system. The three most common solutions are summarised as follows:

— Preprocess the data so that the input language more closely resembles the output language, by means of either enriched input models [1, 38] or segmented translation [35].

— Adapt the language model to make use of the morphological information, i.e. factored models [22].

— Post-process the output of an SMT system to add on the proper inflections by means of morphology generation [37, 6, 16].

In this paper, we further address the introduction of morphology into an SMT system. The main contribution of our work is that we are combining several morphology techniques: preprocessing the data by means of segmented translation and adapting the language model. We are addressing the preprocessing of the data by using the annotations (i.e. tokenisation and Part-of-Speech (POS) Tagger) provided by a Brazilian Portuguese language grammar checker, called COGROO (i.e. Corretor Gramatical para o OpenOffice.org) [33] and the segmentations provided by the MORFESSOR tool [11, 12] which uses unsupervised data-driven methods to divide words into morphemes. For adapting the language model, we are using factored translation models [22]. Then, the two preprocessing and the LM adaptation technique are combined together using the Minimum Bayes Risk MBR strategy [25]. Additionally, we are providing experiments both in an in-domain and out-domain framework. Our morphology work is done specifically for the Brazilian Portuguese and English pair of languages.

3 Statistical machine translation: phrase-based approach

There are several strategies we can follow when translating a pair of languages in SMT. As follows, we briefly describe the phrase-based [24] used in this work.

In general, an SMT system relies on the translation of a source language sentence s into a target language sentence $\hat{t}$. Among all possible target language sentences t we choose the one with the highest probability, as show in equation (1):

$$\begin{aligned} \hat{t} &= \arg\max_t \left[P\left(t|s\right)\right] && (1) \\ &= \arg\max_t \left[P\left(t\right)P\left(s|t\right)\right] && (2) \end{aligned}$$

The probability decomposition shown in equation (2) is based on Bayes' theorem and it is known as the noisy channel approach to SMT [7]. It allows to model independently the target language model $P\left(t\right)$ and the source translation model $P\left(s|t\right)$. The basic idea of this approach is to segment the given source sentence s into segments of one or more words, then each source segment is translated and the target sentence is composed from these segment translations. On the one hand, the translation model weights how likely words in the foreign language are translation of words in the source language. On the other hand, the language model measures the fluency of hypothesis $\hat{t}$. The search process is represented as the $\arg\max$ operation.

The translation model in the phrase-based approach is composed of phrases. A phrase is a pair of m source words and n target words extracted from a parallel sentence that belongs to a bilingual corpus. The parallel sentences have previously been aligned at the word level [8]. Then, given a parallel sentence aligned at the word level, phrases are extracted following the next criteria: we consider the words that are consecutive in both source and target sides and which are consistent with the word alignment. A phrase is consistent with the word alignment if no word inside the phrase is aligned with one word outside the phrase. Finally, phrase translation probabilities are estimated as relative frequencies [40].

The language model assigns a probability to each target sentence. Standard language models are computed following the n-gram strategy, which considers sequences of n words. In order to compute the probability of an n-gram, it is assumed that the probability of observing the *i*th word in the context history of the preceding *i-1* words can be approximated by the probability of observing it in the shortened context history of the preceding *n-1* words. The main problem with this modelling is that it assigns probability zero to strings that have never been seen before. One way to solve this problem is assigning non-zero probabilities to sentences that have never been seen before by means of smoothing techniques [20].

A variation of the noisy channel approach is the log-linear model [28]. It allows using several models or features and to weight them independently as can be seen in equation (3):

$$\hat{t} = \arg\max_t \left[\sum_{m=1}^{M} \lambda_m h_m(s,t)\right] \qquad (3)$$

This equation should be interpreted as a maximum-entropy framework and as a generalisation of equation (2) [40].

Most common additional features that are used in the maximum-entropy framework (in addition to the standard translation and language model) are the lexical models, the word bonus and the reordering model. The lexical models are particularly useful in cases where the translation model may be sparse. For example, for phrases which may have appeared few times the translation model probability may not be well estimated. Then, the lexical models provide a probability among words [8] and they can be computed in both directions source-to-target and target-to-source. The word bonus is used to compensate the language model which benefits shorter outputs. The reordering model is used to provide reordering between phrases. For example, the lexicalised reordering model [36] classifies phrases by the movement they made relative to the previous used phrase, i.e., for each phrase the model learns how likely it is followed by the previous phrase (monotone), swapped with it (swap) or not connected at all (discontinuous).

The different features or models are optimised in the decoder following the minimum error rate procedure [27]. This algorithm searches for weights minimising a given error measure, or, equivalently, maximising a given translation metric. This algorithm enables the weights to be optimised so that the decoder produces the best translations (according to some automatic metric and one or more references) on a development set of parallel sentences.

4 Morphology integration

Our integration of morphology is done using two different approaches: preprocessing and adapting the language model. The former aims at reducing vocabulary and getting a better coverage in translation without the drawback of introducing errors of generation. The latter aims at supporting the most probable POS n-grams in the final translation.

For preprocessing, we need tools to analyse the words and segment them into morphemes. For adapting the language model, we need a tool that provides POS tags. As follows, we technically describe the tools that we are using to perform this analysis, further details on the experimental part are provided later.

4.1 COGROO

In this work, we use the Brazilian Portuguese language grammar checker COGROO[3] tool, which is a recent tool developed at the Universidade de São Paulo [33]. One relevant characteristic of the COGROO tool is that it has a hybrid architecture, mixing rules and statistics. This tool aims to check grammatical errors such as nominal and verbal agreement and other common errors in Brazilian Portuguese language. Some empirical results are shown in previvous publications [32, 19].

We use COGROO to segment some particular words of Brazilian Portuguese language. A complete list of this word segmentation is shown in Table 1.

a + a/as = à/às
a + aquele/aqueles/aquela/aquelas/aquilo = Ã quele/Ã quela/Ã quelas/Ã quilo
a + o/os = ao/aos
a + o/os = ao/aos
com + mim/nós/si/ti/vós/ = comigo/consigo/contigo/convosco
de + aí/alguém = daí/dalguém
de + algum/alguma/alguns/algumas = dalgum/dalguma/dalguns/dalgumas
de + ali/aquém = dali/daquém
de + aquele/aquela/aqueles/aquelas = daquele/daquela/daqueles/ daquelas
de + aqui/aquilo = daqui/daquilo
de + ele/ela/eles/elas = dele/dela/deles/delas
de + entre = dentre
de + esse/essa/esses/essas = desse/dessa/desses/dessas
de + este/esta/estes/estas = deste/desta/destes/destas
de + isso/isot = disso/disto
de + o/a/os/as = do/da/dos/das
de + outrem/outro/outra/outros/outras = doutrem/doutro/doutra/doutros/ doutras
de + um/uma = dum/duma
de + uns/umas = duns/dumas
esse + outro/outra = essoutro/essoutra
este + outro/outra = estoutro/estoutra
ele + o/a/os/as = lho/lha/lhos/lhas
em + algum/alguma/alguns/algumas = nalgum/nalguma/nalguns/ nalgumas
em + aquele/aquela/aqueles/aquelas = naquele/naquela/naqueles/ naquelas
em + aquilo = naquilo
em + ele/ela/eles/elas = nele/nela/neles/nelas
em + esse/essa/esses/essas = nesse/nessa/nesses/nessas
em + este/esta/estes/estas = neste/nesta/nestes/nestas
em + isso/isto = nisso/nisto
em + o/a/os/as = no/na/nos/nas
em + outro/outra/outros/outras = noutro/noutra/noutros/noutras
em + um/uma = num/numa
em + uns/umas = nuns/numas
por + o/a/os/as = pelo/pela/pelos/pelas
para + a/o/as/os = pra/pro/pras/pros

Table 1. COGROO word segmentation

Additionally, COGROO is used to generate the POS tags used for adapting the language model (see subsection 4.3).

[3]Corretor Gramatical para o OpenOffice (Grammar Checker for OpenOffice), http://cogroo.sourceforge.net/

4.2 MORFESSOR

We are using the MORFESSOR tool [11, 12] to segment words into morphemes. The goal of MORFESSOR is to develop unsupervised data-driven methods that discover the regularities behind word forming in natural languages. In particular, this tool focuses on the discovery of morphemes, which are important in automatic generation and recognition of a language, especially in languages in which words may have many different inflected forms.

In particular, we are using the MORFESSOR Categories-MAP model which has a more sophisticated formulation than previous versions [12]. The main difference relies on the fact that it is a complete maximum a posteriori model, which means that it does not need to rely on heuristics in order to determine the optimal size of the morph lexicon. The Categories-ML model introduces a hierarchical lexicon structure: each morph in the lexicon consists either of a string of letters or of two submorphs, which are themselves present in the lexicon. The submorphs can in turn recursively consist of shorter submorphs. Not all morphs in the lexicon need to be *morpheme-like* in the sense that they carry meaning. Some morphs correspond more closely to syllables and other short fragments of words.

The hierarchical structure provides different mechanisms for preventing over- and under-segmentation than the heuristics used in Categories-ML. In a morpheme segmentation task, under-segmentation can be avoided by expanding a lexical item into the submorphs it consists of. In order not to create the opposite problem, over-segmentation, the substructures are only expanded as long as they do not contain non-morphemes.

Further information can be found in [11]. The implementation of the algorithm that we have used is available from the webpage of MORFESSOR Categories-MAP software[4].

[4] http://www.cis.hut.fi/projects/morpho/morfessorcatmapdownloadform.shtml

4.3 Language model adaptation

In order to introduce the language model based on POS tags, we use the factored-based approach. Inspired on the factored-based language models [5], the factored-based approach is an extension of the phrase-based approach presented in Section 3. It adds additional annotation at the word level. A word in this framework is not anymore only a token, but a vector of factors that represent different levels of annotation such as stems and POS.

The translation of factored representations of input words into the factored representations of output words is broken up into a sequence of mapping steps that either translate input factors into output factors, or generate additional output factors from existing output factors.

Factored translation models follow closely the statistical modelling approach of phrase-based models (in fact, phrase-based models are a special case of factored models). The main difference lies in the preparation of the training data and the type of models learned from the data.

5 Evaluation Framework

This section introduces the details of the evaluation framework. We report the translation and the IR system details including corpus statistics, a description of how we built the systems and the evaluation details.

5.1 SMT data

The parallel corpus used to train the SMT system is taken from the Brazilian Portuguese/English bilingual collections of the online issue of the scientific news Brazilian magazine REVISTA PESQUISA FAPESP [2]. See statistics in Table 2. An extra evaluation test (*Eval*) is extracted from a literary collection kindly provided by Stella E. O. Tagnin from the University of São Paulo (USP) hosted by the COMET Project [5]. This *Eval* corpus is used to test the performance of our approaches in an out-domain framework.

[5] http://www.fflch.usp.br/dlm/comet

		PT_{BR}	EN
Train	Sentences	160k	160k
	Words	4,1M	4,3M
	Vocabulary	99,5k	74.7k
Development	Sentences	1375	1375
	Words	34.3k	37.6k
	Vocabulary	6.8k	5.7k
Test	Sentences	1608	1608
	Words	36.8k	38.3k
	Vocabulary	7.3k	6.2k
Eval	Sentences	1600	1600
	Words	29.3k	30.5k
	Vocabulary	9.3k	8.0k

Table 2. Basic characteristics of the SMT experimental dataset.

5.2 Phrase-based and factored-based approaches

Our translation systems were built using MOSES [23]. We used the default MOSES parameters which includes the grow-diagonal-final-and alignment symmetrisation, the lexicalised reordering, relative frequencies, lexical weights and phrase bonus for the translation model (with phrases up to length 10), a 5-gram language model using Kneser-Ney smoothing and a word penalty model built with the SRILM toolkit [34]. Therefore, all these different features are combined in equation (3). The optimisation was done using MERT [27]. For word aligning, we used the standard software GIZA++ [29].

The factored-based translation extension used the same decoder and default parameters as described in the manual webpage[6]. We limited the factor models to the use of POS.

5.3 Morphology segmentation

In this section we report the experimental parameters for the tools that have been used as morphology segmentators. As mentioned, the COGROO tool is used to segment some particular words (i.e. *ao* into *a o*, see Table 1) and provide the text with POS tags. The MORFESSOR tool is used to segment words into morphemes.

[6]http://www.statmt.org/moses/

5.3.1 COGROO - Segmentation and POS tagger training details

For Brazilian Portuguese, the generation of POS tags has been trained on the CETENGOLHA corpus [7]. CETENFOLHA is a 24-million words Brazilian Portuguese POS-tagged corpus, based on journalistic essays. It is not colloquial, generally written in third person. To improve its performance, COGROO requires dealing with abbreviations. An abbreviation dictionary is employed, which contains entries like *sr., tel., apto.*. This dictionary is especially important for the Sentence Boundary Detector and Tokeniser modules. This dictionary was built using Jspell[8]. Other lexical dictionaries are also part of the system, whose construction was based on several other dictionaries freely distributed provided their licenses were compatible with COGROO. This tagger has an accuracy of 0.961.

For English models, COGROO uses the POS tagger available at the well-known tools of OpenNLP [9]. OpenNLP is based on algorithms like Maxent [4] and Perceptron [13].

5.3.2 MORFESSOR - Segmentation details

One of the parameters affecting MORFESSOR segmentation behaviour is the perplexity threshold (PPL), which, roughly speaking, regulates the aggressiveness with which affixes are postulated. We explored lower and higher values than the default value of 10, and found settings of PPL 200 for Brazilian Portuguese to English and PPL 100 for the other direction to be more effective for this MT task. The tendency is that the higher the PPL, the less segmentation and the less reduction of vocabulary.

Figure 2 shows the effect on both vocabulary (on the training set) and BLEU[10] [30] (on the development set) of different MORFESSOR segmentations for a variety of PPL settings on the training set. Table 2 shows the vocabulary of the original unsegmented data for comparison.

We see that the impact of the PPL threshold on translation quality is not very high. However,

[7]Brazilian Portuguese annotated corpus, http://www.linguateca.pt/cetenfolha, last access: 03-2014

[8]Projeto Natura, http://natura.di.uminho.pt/wiki/doku.php?id=ferramentas:jspell, last access: 03-2014

[9]http://opennlp.sourceforge.net/models-1.5/

[10]BLEU stands for Bilingual Evaluation Understudy and it is a standard automatic evaluation metric in MT

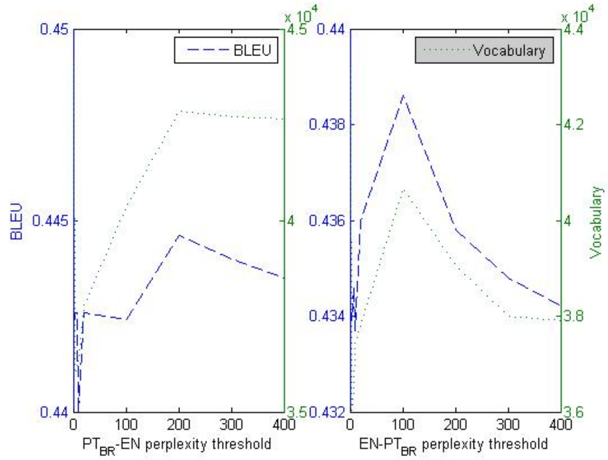

Fig. 2. Translation results for the development set in terms of BLEU and vocabulary size for the training set for different PPL thresholds

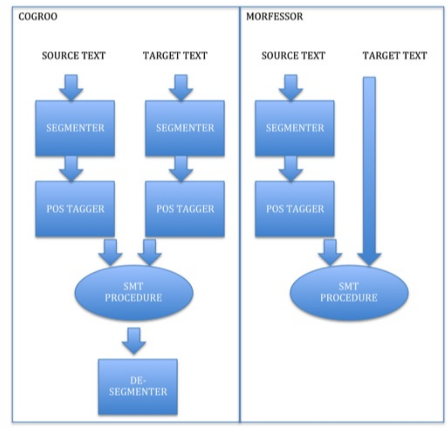

Fig. 3. Pre and post-processing schema when using Cogroo and Morfessor.

we have some interesting variations and we see that the best translation coincides with the highest vocabulary in training, which means lowest segmentation.

5.4 Evaluation and results

Table 3 shows the results in terms of BLEU of the translation system for the in-domain test set. We see that both the COGROO segmentation and MORFESSOR segmentation do not improve the baseline system, whereas the introduction of POS language model improves always its corresponding baseline system (COGROO and MORFESSOR). Note that segmentation when using COGROO is done for both source and target. Therefore, when the target is Brazilian Portuguese, we have to post-process the output to put together segmentations shown in Table 1. However, when segmentation is done using MORFESSOR, it is only done for the source language to avoid post-processing, which would not be error-free. Both schemas are shown in figure 3. When POS are needed, COGROO is used for segmentation and providing POS tags. The improvements over the baseline system are obtained when we combine all systems using MBR. The same conclusions hold for both translation directions.

Significance tests were performed following the “pair bootstrap resampling” method presented in [21], and most of the MBR combination showed better BLEU than the baseline with 99% statistical significance (marked with * in Tables 3 and 4).

Test	PT$_{BR}$-EN	EN-PT$_{BR}$
Baseline	0.3571	0.2426
COGROO	0.3565	0.2375
COGROO+LM$_{POS}$	0.3579	0.2399
MORFESSOR (PPL=200/100)	0.3470	0.2391
MORFESSOR (PPL=200/100)+LM$_{POS}$	0.3491	0.2392
MBR	**0.3623***	**0.2430**

Table 3. Translation results in terms of BLEU. Best results in bold, of which statistically significant improvements marked with (*).

Table 4 shows the results in terms of BLEU of the translation system for the out-of-domain test set. Results are consistent with the ones obtained in the in-domain test set.

6 Case Study: The OnAir system

This case study proposes a multilingual extension for ONAIR which is an ontology-aided IR system applied to retrieve clips from a video collection, described in detail in previous studies [31]. The multilingual extension basically involves allowing the user to search and retrieve either in Brazilian Portuguese or English. In order to perform query translation we use the SMT approach enhanced with morphology as presented in the sections above. Our experiments show that the multilingual system is capable of achieving almost the same quality of that obtained by the monolingual system.

Test	PT_{BR}-EN	EN-PT_{BR}
Baseline	0.1298	0.0694
COGROO	0.1285	0.0647
COGROO+LM_{POS}	0.1317	0.0680
MORFESSOR (PPL=200/100)	0.1233	0.0651
MORFESSOR (PPL=200/100)+LM_{POS}	0.1236	0.0676
MBR	**0.1326***	**0.0701***

Table 4. Out-of-domain translation results in terms of BLEU. Best results in bold, of which statistically significant improvements marked with (*).

6.1 Information Retrieval

ONAIR relies on the vector space model [3] for information retrieval. It was built to receive videos and keywords or their transcriptions, with timeline markers, as input, and to allow the users to query for video excerpts using natural language. When a user query is presented, ONAIR returns a list of video excerpts that best answer the user query.

The video transcriptions are pre-processed, using traditional IR techniques: stemming and stop-word removal, then the vector space model is used for indexing and retrieving. As usual in traditional IR systems, some additional techniques are needed to avoid natural language difficulties like polysemy and synonymy.

6.2 Ontology description

Ontologies are defined in general as an explicit specification for a conceptualisation [17]. As mainly used for IR it can be seen as a set of concepts related by hierarchies and other kind of properties in a specific domain [14]. Ontologies have been commonly used in IR through query expansion and conceptual distance measures [31].

A domain ontology related to the topics from the videos is needed to be able to do the query expansion. By definition, query expansion is the process of reformulating a seed query to improve retrieval performance in IR operations. In particular, the domain ontology is used to measure the conceptual distance among seed query terms and new ones.

6.3 Cross-language extension

The multilingual extension of ONAIR is basically a challenge of cross-language information retrieval (CLIR). Given a query in a source language, the aim of CLIR is retrieving related documents in a target language. [26] identified four types of strategies for matching a query with a set of documents in the context of CLIR by: cognate matching, document translation, query translation or interlingua techniques. From these techniques the most used are the query translation techniques. Query translation methods translate user queries to the language that the documents are written. It is the most popular approach in CLIR experimental systems due to its tractability and convenience. CLIR through query translation methods has been mainly faced by using dictionary-based (i.e. using machine-readable dictionaries, MRD), MT and/or parallel texts techniques [9].

In this case, we are using one of the most popular approaches nowadays which is the standard phrase-based SMT approach, as described in the above sections. Additionally, we are comparing the performance of a standard phrase-based SMT system and a phrase-based SMT system which uses morphology.

6.4 Use case experiments

In this subsection we report the experiments using ONAIR. As follows we describe the data used and discuss the results obtained in monolingual IR and CLIR contexts.

6.4.1 IR data

For testing the IR system in Brazilian Portuguese we used a video collection compiled from interviews with Ana Teixeira, a Brazilian artist. The interviews were made by Paula P. Braga, domain expert and there have been used in previous studies as [31]. The interview was developed in the domain of contemporary art and the system uses a domain ontology to expand queries with related terms. To test the system, a battery of queries was synthesised both for English and Brazilian Portuguese. Statistics of these queries and the corresponding documents for retrieving are shown in Table 5.

		PT-BR	EN
Query	Number	50	50
	Words	349	435
	Vocabulary	155	145
Documents	Number	48	-
	Words	8.2k	-
	Vocabulary	2.4k	-

Table 5. Basic characteristics of the query and documents dataset for the Ana Teixerira videos.

6.4.2 Comparing IR and CLIR system's performance

We performed the following experiments: two experiments using a monolingual IR, recovered from previous publications [31], and one using a CLIR system, similarly to previous publications [10] but with the extension of adding morphology knowledge in MT. We describe the corresponding systems as follows:

1. IR system: the original system analysed was the system described in subsection 6.1, with the following configuration: *mono-kw-fulltext* which uses the results of retrieval using keywords and transcriptions, the best configuration for ONAIR as described in [31]

2. CLIR system (*smt-baseline-kw-fulltext*): this system is the concatenation of the baseline SMT system from Section (5) and the IR system from the point above in this list. The performance of the SMT system in terms of BLEU is 0.0977.

3. CLIR system (*smt-mbr-kw-fulltext*): this system is the concatenation of the SMT system improved with morphology (MBR) from Section (5) and the IR system from the point above in this list. The performance of the SMT system in terms of BLEU is 0.1348, note that this represents an improvement of almost 4 points BLEU of the corresponding baseline.

Figure 4 shows the results of the f-measure run over the 50 queries analysed in our experiments in the three configurations presented above and the BLEU measure for the translation of each query (for the best SMT system).

We have computed the Pearson correlation[11] among BLEU and f-measure and we found out that

[11] http://mathworld.wolfram.com/CorrelationCoefficient.html

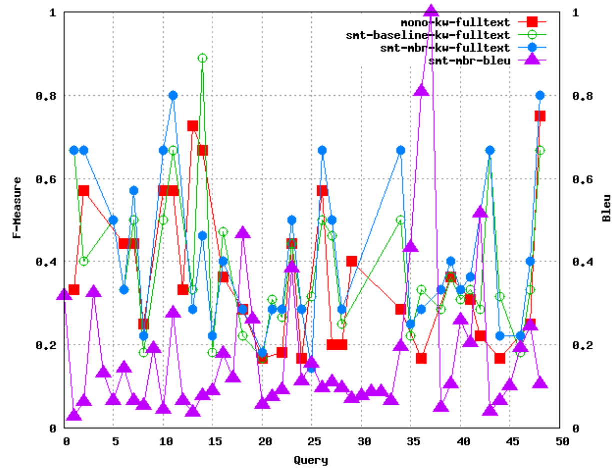

Fig. 4. F-measure for the systems analysed and BLEU for the best SMT system.

it is of 7.73%; among BLEU and precision is of 19.46% and among BLEU and recall is of -2.23%. So, the quality of MT (in terms of BLEU) is not related to the quality of information retrieval (IR) (in terms of f-measure, precision and recall).

Surprisingly, experiments show that the CLIR system, for specific queries, is capable of outperforming the IR system. For these queries, the translation system uses a more adequate word, which means that it would be possible to use MT to perform query expansion. It would be interesting to built the CLIR system with the *n*-best translations.

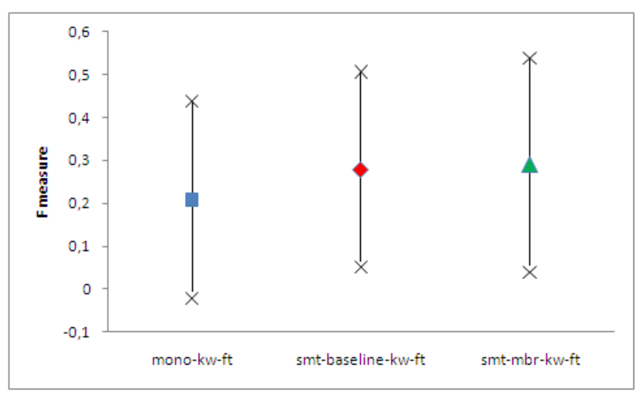

Fig. 5. Average f-measure for the systems analysed.

Figure 5 shows the f-measure in average for all systems that we experimented. Here, we observe that the f-measure of with respect to the CLIR system (*smt-baseline-kw-fulltext*) is slightly better than its comparable IR system (*mono-kw-fulltext*). And the best retrieval system is *smt-mbr-kw-fulltext*.

Finally, Figure 1 shows some translation examples. It shows the input to the CLIR system: *smt-mbr-kw-fulltext*, the corresponding translation output and the corresponding reference (i.e. the input of the IR system). The two first examples report cases where the CLIR system performs worse than the IR system (*mono-kw-fulltext*) in terms of f-measure. The second two examples report cases where the CLIR system performs better than the IR system in terms of f-measure. Coherently, in the first case, the translation shows a poorer quality than in the second case.

INPUT: How did you become an artist? MBR: Como o senhore se um artista? REFERENCE: Como você virou artista
INPUT: Do you make only interventions or also paintings? MBR: O senhor faz apenas intervenções ou também pinturas? REFERENCE: Você só faz intervenções ou faz também pintura?
INPUT: I loved his work. MBR: Eu adorava sua obra. REFERENCE: Adorei seu trabalho.
INPUT: Have you ever exposed abroad? MBR: O senhor já exposta no exterior? REFERENCE: Você já expôs no exterior?

Fig. 6. Translation examples.

7 Conclusions

This paper works in enhancing MT by introducing morphology knowledge into a standard system. In addition, we see the impact of that improvement into a CLIR on-line application. As follows we detail the contributions of this paper:

1. Description of two approaches and their combination to integrate morphology into a standard phrase-based approach. Firstly, we have used specific tools to segment the input into morphemes. Secondly, we have introduced a language model with POS information. Both approaches are successfully combined using the MBR approach. Consistent and significant improvements are reported in in-domain and out-of-domain evaluation sets and over two translation directions: from Brazilian Portuguese into English and the other way round.

2. Experimentation with sophisticated tools to segment the input into morphemes. These tools are COGROO, and MORFESSOR. This is the first work in SMT which uses COGROO and it has been shown useful for introducing POS tags.

3. Preparation and compilation of new data sets in the pair Brazilian Portuguese/English. These data sets are parallel corpus at the level of sentence.

4. Case study that generates a cross-language extension for the ONAIR system, which is in essence an IR system using ontologies to expand queries. The cross-language extension has been done using a state-of-the-art SMT system with or without morphology. Experiments show that the best configuration for the CLIR system (including morphology in the SMT system) can get competitive results compared to the IR system.

As further work, we want to explore different linguistic and statistical techniques (focusing on semantics) to be introduced in the SMT system in order to improve the translation of queries which are essentially out-of-domain.

Acknowledgements

The author would like to specially thank Prof. Renata Wassermann for her help and assistance; Christian Paz-Trillo for his support on the ONAIR system; William Colen for his help with the COGROO system; Stella O. Tagnin for providing the out-of-domain corpus and Fabiano Luz for his dedication to parallelize this corpus.

This work has been partially supported by FAPESP through the ONAIR project (2010/19111-9) and the *visiting researcher program* (2012/02131-2), by the Spanish Ministry of Economy and Competitiveness through the Juan de la Cierva fellowship program and the contract TEC2012-38939-C03-02, as well as from the Seventh Framework Program of the European Commission through the International Outgoing Fellowship Marie Curie Action (IMTraP-2011-29951).

References

1. **Avramidis, E. & Koehn, P.** (**2008**). Enriching morphologically poor languages for statistical machine translation. In *Proc. of the conference of the Association for Computational Linguistics and Human Language Technology (ACL-HLT)*. 763–770.
2. **Aziz, W. & Specia, L.** (**2011**). Fully automatic compilation of a Portuguese-English parallel corpus for statistical machine translation. In *STIL 2011*. Cuiabá, MT.
3. **Baeza-Yates, R. & Ribeiro-Neto, B.** (**1999**). *Modern Information Retrieval*. Addison Wesley Longman.
4. **Berger, A., Pietra, S. D., & Pietra., V. D.** (**1996**). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1), 39–71.
5. **Bilmes, J. A. & Kirchhoff, K.** (**2003**). Factored language models and generalized parallel backoff. In *Proceedings of HLT/NACCL*. 4–6.
6. **Bojar, O. & Tamchyna, A.** (**2011**). Forms wanted: Training smt on monolingual data. In *Workshop of Machine Translation and Morphologically-Rich Languages*.
7. **Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L., & Roossin, P. S.** (**1990**). A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2), 79–85.
8. **Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L.** (**1993**). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2), 263–311.
9. **Chen, J. & Bao, Y.** (**2009**). Cross-language search: The case of google language tools. *First Monday*, 14(3-2).
10. **Costa-jussà, M. R., Paz-Trillo, C., & Wassermann, R.** (**2012**). Initial approaches on cross-lingual informaiton retrieval using statistical machine translation on user queries. In *Proceedings of the Joint V Seminar on Ontology Research in Brazil and VII International Workshop on Metamodels, Ontologies and Semantic Technologies*. Recife, Brazil, 25–35.
11. **Creutz, M. & Lagus, K.** (**2005**). Inducing the morphological lexicon of a natural language from unannotated text. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*. Espoo, Finland.
12. **Creutz, M., Lagus, K., & Virpioja, S.** (**2005**). Unsupervised morphology induction using morfessor. In *Finite-State Methods and Natural Language Processing, 5th International Workshop, FSMNLP 2005, Helsinki, Finland, September 1-2, 2005. Revised Papers*, volume 4002 of *Lecture Notes in Computer Science*. Springer. ISBN 3-540-35467-0, 300–301.
13. **Daume, H. C., III** (**2006**). *Practical structured learning techniques for natural language processing*. Ph.D. thesis, Los Angeles, CA, USA. AAI3337548.
14. **Ding, Y.** (**2001**). Ir and ai: The role of ontology. In *International Conference of Asian Digital Libraries*.
15. **Forcada, M. L.** (**2006**). Open-source machine translation: an opportunity for minor languages. In *Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages)*.
16. **Formiga, L., Costa-jussà, M. R., Mariño, J. B., Fonollosa, J. A. R., Barrón-Cedeño, A., & Márquez, L.** (**2013**). The TALP-UPC phrase-based translation systems for WMT13: System combination with morphology generation, domain adaptation and corpus filtering. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria, 134–140.
17. **Gruber, T. R.** (**1993**). A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2), 199–220.
18. **Karageorgakis, P., Potamianos, A., & K., I.** (**2005**). Towards incorporating language morphology into statistical machine translation systems. In *Automatic Speech Recognition and Understanding Workshop*.
19. **Kinoshita, J., Salvador, L. N., & Menezes, C. E.** (**2007**). Cogroo - an openoffice grammar checker. In *Proceedings of the Seventh international Conference on intelligent Systems Design and Applications (ISDA)*. IEEE Computer Society, 525–530.
20. **Kneser & Ney** (**1995**). Improved backing-off for m-gram language modeling. In *IEEE Inte. Conf. on Acoustics, Speech and Signal Processing*. Detroit, MI, 49–52.
21. **Koehn, P.** (**2004**). Statistical Significance Tests For Machine Translation Evaluation. In *Proceedings of EMNLP*, volume 4. 388–395.
22. **Koehn, P. & Hoang, H.** (**2007**). Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Association for Computational Linguistics, Prague, Czech Republic, 868–876.

23. **Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E.** (**2007**). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*. Prague, Czech Republic, 177–180.

24. **Koehn, P., Och, F., & Marcu, D.** (**2003**). Statistical Phrase-Based Translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics*.

25. **Kumar, S. & Byrne, W.** (**2002**). Minimum bayes-risk word alignments of bilingual texts. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02. Stroudsburg, PA, USA, 140–147.

26. **Oard, D. W. & Diekema, A. R.** (**1998**). Cross-Language information retrieval. *Annual Review of Information Science and Technology (ARIST)*, 33, 223–256.

27. **Och, F.** (**2003**). Minimum Error Rate Training In Statistical Machine Translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics*. 160–167.

28. **Och, F. & Ney, H.** (**2002**). Dicriminative training and maximum entropy models for statistical machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA, 295–302.

29. **Och, F. J. & Ney, H.** (**2000**). Improved Statistical Alignment Models. In *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics*. Hongkong, China, 440–447.

30. **Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J.** (**2002**). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02. 311–318.

31. **Paz-Trillo, C., Wassermann, R., & Braga, P. P.** (**2005**). An information retrieval application using ontologies. *J. Braz. Comp. Soc.*, 11(2), 17–31.

32. **Silva, W., Finger, M., & Menezes, C.** (**2010**). Open text annotators using apache uima. In *PROPOR*.

33. **Silva, W. D.** (**2012**). *CoGrOO: Corretor Gramatical acoplável ao LibreOffice e Apache OpenOffice*. CCSL IME/USP, São Paulo, Brasil.

34. **Stolcke, A.** (**2002**). SRILM: an extensible language modeling toolkit. In *Proc. of the Int. Conf. on Spoken Language Processing*. Denver, CO, 901–904.

35. **S.Virpioja, Väyrynen, J., Creutz, M., & Sadeniemi, M.** (**2007**). Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *Machine Translation Summit XI*. 491–498.

36. **Tillman, C.** (**2004**). A Block Orientation Model for Statistical Machine Translation. In *HLT-NAACL*.

37. **Toutanova, K., Suzuki, H., & Ruopp, A.** (**2008**). Applying morphology generation models to machine translation. In *Proc. of the conference of the Association for Computational Linguistics and Human Language Technology (ACL-HLT)*. Columbus, Ohio, 514–522.

38. **Ueffing, N. & Ney, H.** (**2003**). Using pos information for statistical machine translation into morphologically rich languages. In *Proc. of the 10th conference on European chapter of the Association for Computational Linguistics (EACL)*. Stroudsburg, PA, USA, 347–354.

39. **Way, A. & Gough, N.** (**2005**). Comparing example-based and statistical machine translation. *Natural Language Engineering*, 11(3), 295–309.

40. **Zens, R., Och, F., & Ney, H.** (**2002**). Phrase-based statistical machine translation. In **Verlag, S.**, editor, *Proc. German Conference on Artificial Intelligence (KI)*.

Article received on 06/12/2012; accepted on 16/01/2013.

BASELINE: laboratório finaliza projeto de um novo anel para produção de lu z síncrotron

SEGMENTED PPL=10: labor +a +tório final +iza pro+ jeto de um novo anel par a pro+ du+ ção de luz síncro tron

SEGMENTED PPL=100: labora tório final iza projeto de um novo anel para pro du +ção de luz sín cro tron

SEGMENTED PPL=200:labora tório final iza projeto de um novo anel para prod u ção de luz sín crotron

SEGMENTED PPL=400:labora tório finaliza projeto de um novo anel para produ ção de luz sín crotron

---

BASELINE: inclusive , nosso setor de importação do icb tem conseguido importar estes camundongos .

SEGMENTED PPL=10:inclusive , nosso setor de import +a +ção do icb tem con+ segui +do import +a +r estes camundongo +s .

SEGMENTED PPL=100: inclusive , nosso setor de importa +ção do icb tem con+ segui +do importa +r estes camundongo +s .

SEGMENTED PPL=200:inclusive , nosso setor de importa ção do icb tem con segui +do importa +r estes camundongo +s .

SEGMENTED PPL=400:inclusive , nosso setor de importa ção do icb tem con segui +do importa +r este +s camundongo +s .

BASELINE:laboratory completes project for new synchrotron light production ring

SEGMENTED PPL=10: labor +ator +y complete +s project for new synchrotron light pro+ duct +ion ring

SEGMENTED PPL=100:labor atory complet +e +s project for new synchrotron light production ring

SEGMENTED PPL=200:labor atory complete +s project for new synchrotron light product ion ring

SEGMENTED PPL=400:labor atory complete +s project for new synchrotron light production ring

---

BASELINE: in fact , our icb import sector has managed to import these mice .

SEGMENTED PPL=10:in fact , our icb import sector has manage +d to import these mice .

SEGMENTED PPL=100:in fact , our icb import sector has managed to import the+ se mice .

SEGMENTED PPL=200:in fact , our icb import sector has managed to import these mice .

SEGMENTED PPL=400:in fact , our icb import sector has managed to import these mice

Fig. 1. Segmentation examples for different perplexity thresholds, in Brazilian-Portuguese (above) and English (down). Symbol + at the beginning of the word indicates a prefix, at the end of the word indicates a suffix.